

Express Mail Mailing Label No. EL318654616US

APPLICATION

OF

Samuel Keith McDowell

FOR

UNITED STATES LETTERS PATENT

ON

SYSTEM AND METHOD FOR PROVIDING INTERACTIVE AUDIO IN A
MULTI-CHANNEL AUDIO ENVIRONMENT

Docket No. 262-23-219

Assigned to

Digital Theater Systems (DTS), Inc.

SYSTEM AND METHOD FOR PROVIDING INTERACTIVE AUDIO IN A MULTI-CHANNEL AUDIO ENVIRONMENT

FIELD OF THE INVENTION

5 This invention relates to fully interactive audio systems and more specifically to a system and method of rendering real-time multi-channel interactive digital audio to create a rich immersive surround sound environment suitable for 3D gaming, virtual reality and other interactive audio applications.

BACKGROUND OF THE INVENTION

Recent developments in audio technology have focused on creating real-time interactive positioning of sounds anywhere in the three-dimensional space surrounding a listener (the “sound field”). True interactive audio provides not only for the ability to create the sound on-demand but the ability to position that sound precisely in the sound field. Support for such technologies can be found in a variety of products but most frequently in software for video games to create a natural, immersive, and interactive audio environments. Applications extend beyond gaming into the entertainment world in the form of audiovisual products such as DVD, and also into video-conferencing, simulation systems and other interactive environments.

20 Advances in audio technology have proceeded in the direction of making the audio environment “real” to a listener. Surround sound developments followed, first in the analog domain with the development of HRTFs, Dolby Surround and later in the digital domain with AC-3, MPEG, and DTS to immerse a listener in the surround sound environment.

25 In order to portray a realistic synthetic environment, virtual sound systems use binaural technology and psychoacoustic cues to create the illusion of surround audio without the need for multiple speakers. The majority of these virtualized 3D audio technologies are based on the concept of HRTFs (Head-Related Transfer Functions). The original digitized sound is convolved in real-time with the left- and right-ear HRTFs corresponding to the desired spatial location, right- and left-ear binaural signals are produced, which when heard seem to come from the desired location. To position the

sound, the HRTFs are changed to those for the desired new location and the process repeated. A listener can experience nearly free-field listening through headphones if the audio signals are filtered with that listener's own HRTFs. However, this is often impractical and experimenters have searched for a set of general HRTFs that have good 5 performance for a wide range of listeners. This has been difficult to accomplish with a specific obstacle being front-back confusion, which describes the sensation that sounds either in front of or behind the head are coming from the same direction. Despite its drawbacks, HRTF methods have been successfully applied to both PCM audio and with much lessened computational load to compressed MPEG audio. Although virtual 10 surround technologies based on HRTFs provide significant benefits in situations where full home theater set-ups are not practical these current solutions do not provide any means for interactive positioning of specific sounds.

The Dolby Surround system is another method to implement positional audio. Dolby Surround is a matrix process that enables a stereo (two-channel) medium to carry 15 four-channel audio. The system takes four-channel audio and generates two channels of Dolby Surround encoded material identified as left total (Lt) and right total (Rt). The encoded material is decoded by a Dolby Pro-Logic decoder producing a four-channel output; a left channel, a right channel, a center channel and a mono surround channel. The center channel is designed to anchor voices at the screen. The left and right channels 20 are intended for music and some sound effects, with the surround channel primarily dedicated to the sound effects. The surround sound tracks are pre-encoded in Dolby Surround format, and thus they are best suited for movies, and are not particularly useful in interactive applications such as video games. PCM audio can be overlaid on the Dolby Surround sound audio to provide a less controllable interactive audio experience. 25 Unfortunately, mixing PCM with Dolby Surround Sound is content dependant and overlaying PCM audio on the Dolby Surround sound audio tends to confuse the Dolby Prologic decoder, which can create undesirable surround artifacts and crosstalk.

To improve channel separation digital surround sound technologies such as Dolby 30 Digital and DTS provide six discrete channels of digital sound in a left, center and right front speakers along with separate left surround and right surround rear speakers and a subwoofer. Digital surround is a pre-recorded technology and thus best suited for movies

and home A/V systems where the decoding latency can be accommodated and in its present form it is not particularly useful for interactive applications such as video games. However, since Dolby Digital and DTS provide high fidelity positional audio, have a large installed base of home theater decoders, definitions for a multi-channel 5.1 speaker
5 format and product available for market, they present a highly desirable multi-channel environment for PCs and in particular console based gaming systems if they could be made fully interactive. However, the PC architecture has generally been unable to deliver multi-channel digital PCM audio to home entertainment systems. This is primarily due to the fact that the standard PC digital output is through a stereo based
10 S/PDIF digital output connector.

Cambridge SoundWorks offers a hybrid digital surround/PCM approach in the form of the DeskTop Theater 5.1 DTT2500. This product features a built-in Dolby Digital decoder that combines pre-encoded Dolby Digital 5.1 background material with interactive four-channel digital PCM audio. This system requires two separate
15 connectors; one to deliver the Dolby Digital and one to deliver the 4-channel digital audio. Although a step forward, Desktop Theater is not compatible with the existing installed base of Dolby Digital decoders and requires sound cards supporting multiple channels of PCM output. The sounds are reproduced from the speakers located at known locations, but the goal in an interactive 3D sound field is to create a believable
20 environment in which sounds appear to originate from any chosen direction about the listener. The richness of the DeskTop Theater's interactive audio is further limited by the computation requirements needed to process the PCM data. Sideways localization, which is a critical component of a positional audio environment is computationally expensive to apply on time-domain data, as are the operations of filtering and
25 equalization.

The gaming industry needs a low cost fully-interactive low latency immersive digital surround sound environment suitable for 3D gaming and other interactive audio applications that allows the gaming programmer to mix a large number of audio sources and to precisely position them in the sound field and which is compatible with the
30 existing infrastructure of home theater Digital Surround Sound systems.

SUMMARY OF THE INVENTION

In view of the above problems, the present invention provides a low cost fully interactive immersive digital surround sound environment suitable for 3D gaming and other high fidelity audio applications, which can be configured to maintain compatibility 5 with the existing infrastructure of Digital Surround Sound decoders.

This is accomplished by storing each audio component in a compressed format that sacrifices coding and storage efficiency in favor of computational simplicity, mixing the components in the subband domain rather than the time domain, recompressing and packing the multi-channel mixed audio into the compressed format and passing it to a 10 downstream surround sound processor for decoding and distribution. As the multi-channel data is in a compressed format, it can be passed across a stereo based S/PDIF digital output connector. Techniques are also provided for “looping” compressed audio, which is an important and standard feature in gaming applications that manipulate PCM 15 audio. In addition, decoder sync is ensured by transmitting frames of “silence” whenever mixed audio is not present either due to processing latency or the gaming application.

More specifically, the components are preferably encoded into a subband representation, compressed and packed into a data frame in which only the scale factors and subband data change from frame-to-frame. This compressed format requires significantly less memory than standard PCM audio but more than that required by 20 variable length code storage such as used in Dolby AC-3 or MPEG. More significantly this approach greatly simplifies the unpack/pack, mix and decompress/compress operations thereby reducing processor utilization. In addition, fixed length codes (FLCs) aid the random access navigation through an encoded bitstream. High levels of throughput can be achieved by using a single predefined bit allocation table to encode the 25 source audio and the mixed output channels. In the currently preferred embodiment, the audio renderer is hardcoded for a fixed header and bit allocation table so that the audio renderer only need process the scale factors and subband data.

Mixing is achieved by partially decoding (decompressing) only the subband data from components that are considered audible and mixing them in the subband domain. 30 The subband representation lends itself to a simplified psychoacoustic masking technique so that a large number of sources can be rendered without increasing processing

complexity or reducing the quality of the mixed signal. In addition, since multi-channel signals are encoded into their compressed format prior to transmission, a rich high-fidelity unified surround sound signal can be delivered to the decoder over a single connection.

5 These and other features and advantages of the invention will be apparent to those skilled in the art from the following detailed description of preferred embodiments, taken together with the accompanying drawings, in which:

BRIEF DESCRIPTION OF THE DRAWINGS

10 Figs. 1a through 1c are block diagrams of different gaming configurations in accordance with the present invention;

Fig. 2 is a block diagram of the application layer structure for a fully interactive surround sound environment;

Fig. 3-1 & 3-2 (collectively Fig. 3) are a flowchart of the audio rendering layer shown in
15 Fig 2;

Fig. 4 is a block diagram of the packing process for assembling and queuing up the output data frames for transmission to a surround sound decoder;

Fig. 5 is a flow chart illustrating the looping of compressed audio;

Fig. 6 is a diagram depicting the organization of the data frames;

20 Fig. 7 is a diagram depicting the organization of the quantized subband data, scale factors and bit allocation in each frame;

Fig. 8 is a block diagram of the subband domain mixing process;

Fig. 9 is a diagram that illustrates the psychoacoustic masking effects;

Figs. 10a through 10c diagram the bit extraction process for packing and unpacking each
25 frame; and

Fig. 11 is diagram that illustrates the mixing of the specified subband data.

DETAILED DESCRIPTION OF THE INVENTION

DTS Interactive provides a low cost fully interactive immersive digital surround sound environment suitable for 3D gaming and other high fidelity audio applications. DTS Interactive stores the component audio in a compressed and packed format, mixes

the source audio in the subband domain, recompresses and packs the multi-channel mixed audio into the compressed format and passes it to a downstream surround sound processor for decoding and distribution. As the multi-channel data is in a compressed format, it can be passed across a stereo based S/PDIF digital out put connector. DTS

5 Interactive greatly increases the number of audio sources that can be rendered together in an immersive multi-channel environment without increasing the computational load or degrading the rendered audio. DTS Interactive simplifies equalization and phase positioning operations. In addition techniques are provided for “looping” compressed audio and decoder sync is ensured by transmitting frames of “silence” whenever source

10 audio is not present where silence includes true silence or low level noise. DTS Interactive is designed to maintain backward compatibility with the existing infrastructure of DTS Surround Sound decoders. However, the described formatting and mixing techniques could be used to design a dedicated gaming console that would not be limited to maintaining source and/or destination compatibility with the existing decoder.

15

DTS Interactive

The DTS Interactive system is supported by multiple platforms, of which there are the DTS 5.1 multi-channel home theatre system 10, which includes a decoder and an AV amplifier, a sound card 12 equipped with hardware DTS decoder chipset with an AV

20 amplifier 14, or a software implemented DTS decoder 16 with an audio card 18 and an AV amplifier 20, see figures 1a, 1b and 1c. All those systems require a set of speakers named left 22, right 24, left surround 26, right surround 28, center 30 and sub-woofer 32, a multi-channel decoder and a multi-channel amplifier. The decoder provides digital S/PDIF or other input for supplying compressed audio data. The amplifier powers six

25 discrete speakers. Video is rendered on a display or projection device 34, usually a TV or other monitor. A user interacts with the AV environment through a human interface device (HID) such as a keyboard 36, mouse 38, position sensor, trackball or joy stick.

Application Programming Interface (API)

30 As shown in figures 2 and 3, the DTS Interactive system consists of three layers: the application 40, the application programming interface (API) 42 and the audio

renderer 44. The software application could be a game, or maybe a music playback/composition program, which takes component audio files 46 and assigns to each some default positional character 48. The application also accepts interactive data from the user via an HID 36/38.

5 For each game level, frequently used audio components are loaded into memory (step 50). Because each component is treated as an object the programmer is kept unaware of the sound format and rendering details, he need only be concerned with the absolute position to the listener and the effects processing that might be desired. The DTS Interactive format allows these components to be mono, stereo or multi-channel with or
10 without low frequency effects (LFE). Since DTS Interactive stores the components in a compressed format (see figure 6) valuable system memory is saved that can otherwise be used for higher resolution video rendering, more colors, or more textures. The reduced file size resulting from the compressed format also permits rapid on demand loading from the storage media. The sound components are provisioned with parameters to detail the
15 position, equalization, volume and necessary effects. These details will influence the outcome of the rendering process.

API layer 42 provides an interface for the programmer to create and control each sound effect and also provides isolation from the complicated real-time audio rendering process that deals with the mixing of the audio data. Object orientated classes create and
20 control the sound generation. There are several class members at the programmers disposal, which are as follows: load, unload, play, pause, stop, looping, delay, volume, equalization, 3D position, maximum and minimum sound dimensions of the environment, memory allocation, memory locking and synchronization.

The API generates a record of all sound objects created and loaded into memory
25 or accessed from media (step 52). This data is stored in an object list table. The object list does not contain the actual audio data but rather tracks information important to the generation of the sound such as information to indicate the position of the data pointer within the compressed audio stream, the position coordinates of the sound, the bearing and distance to the listener's location, the status of the sound generation and any special
30 processing requirements for mixing the data. When the API is called to create a sound object, a reference pointer to the object is automatically entered into the object list. When

an object is deleted, the corresponding pointer entry in the object list is set to null. If the object list is full then a simple age based caching system can choose to overwrite old instances. The object list forms the bridge between the asynchronous application, the synchronous mixer and compressed audio generator processes.

5 The classes inherited by each object permit start, stop, pause, load and unload functions to control the generation of the sound. These controls allow the play list manager to examine the object list and construct a play list 53 of only those sounds that are actively playing at that moment in time. The manager can decide to omit a sound from the play list if it is paused, stopped, has completed playing or has not been delayed
 10 sufficiently to commence playing. Each entry in the play list is a pointer to individual frames within a sound that must be examined and if necessary piecewise unpacked prior to mixing. Since frame sizes are constant, manipulation of the pointer permits playback positioning, looping and delay of the output sound. This pointer value indicates the current decoding position within the compressed audio stream.

15 The positional localization of sounds requires the assignment of sounds to individual rendering pipelines or execute buffers that in turn map directly onto the arrangement of the loudspeakers (step 54). This is the purpose of the mapping function. Position data for entries in the frame list are examined to determine which signal processing functions to apply, renew the bearings and direction of each sound to the
 20 listener, alter each sound depending on physical models for the environment, determine mixing coefficients and allocate audio streams to available and most appropriate speakers. All parameters and model data are combined to deduce modifications to the scale factors associated with each compressed audio frame entering a pipeline. If side localization is desired, data from the phase shift tables are indicated and indexed.

25

Audio Rendering

As shown in figures 2 and 3, audio rendering layer 44 is responsible for mixing the desired subband data 55 according to the 3D parameters 57 set by the object classes. The mixing of multiple audio components requires the selective unpacking and
 30 decompression of each component, summing of correlated samples and the calculation of a new scale factor for each subband. All processes in the rendering layer must function in

real-time to deliver a smooth and continuous flow of compressed audio data to the decoding system. A pipeline receives a listing of the sound objects in play and, from within each object, directions for the modification of the sound. Each pipeline is designed to manipulate the component audio according to the mixing coefficients and to mix an output stream for a single speaker channel. The output streams are packed and multiplexed into a unified output bitstream.

More specifically, the rendering process commences by unpacking and decompressing each component's scale factors into memory on a frame-by-frame basis (step 56), or alternately multiple frames at a time (see figure 7). At this stage only the scale factor information for each subband is required to assess if that component, or portions of the component, will be audible in the rendered stream. Since fixed length coding is used, it is possible to unpack and decompress only that part of the frame that contains the scale factors thereby reducing processor utilization. For SIMD performance reasons each 7-bit scale factor value is stored as a byte in memory space, and aligned to a 32-byte address boundary to ensure that a cache line read will obtain all scale factors in one cache fill operation and not cause cache memory pollution. To further speed this operation, the scale factors may be stored as bytes in the source material and organized to occur in memory on 32 byte address boundaries.

The 3D parameters 57 provided by the 3D position, volume, mixing and equalization are combined to determine a modification array for each subband that is used to modify the extracted scale factors (step 58). Because each component is represented in the subband domain equalization is a trivial operation of adjusting the subband coefficients as desired via the scale factors.

In step 60, the maximum scale factors indexed for all elements in the pipeline are located and stored to an output array, which is suitably aligned in memory space. This information is used to decide the need to mix certain subband components.

At this point, step 62, masking comparisons are made with the other pipelined sound objects to remove the inaudible subbands from the speaker pipelines (see figures 8 and 9 for details). The masking comparisons are preferably done for each subband independently to improve speed and are based upon the scale factors for the objects referenced by the list. A pipeline contains only that information which is audible from a

single speaker. If an output scale factor is below the threshold of human hearing then the output scale factor may be set to zero and in doing so remove the need to mix the corresponding subband components. The advantage of DTS Interactive over manipulation of PCM time-domain audio is that the gaming programmer is allowed to
5 use many more components and rely on the masking routine to extract and mix only the audible sounds at any given time without excess computations.

Once the desire subbands are identified, the audio frames are further unpacked and decompressed to extract only the audible subband data (step 64), which is stored as left shifted DWORD format in memory (see figures 10a-10c). Throughout the
10 description the DWORD is assumed without loss of generality to be 32 bits. In the gaming environment, the price paid in lost compression for using FLCs is more than compensated by the reduction in the number of computations required to unpack and decompress the subband data. This process is further simplified by using a single predefined bit allocation table for all of the components and channels. FLCs enable
15 random positioning of the read position at any subband within the component.

In step 66, phase positioning filtering is applied to the subband data for bands 1 and 2. The filter has specific phase characteristics and need only be applied over the frequency range 200Hz to 1200Hz where the ear is most sensitive to positional cues. Since phase position calculations are only applied to first two bands of the 32 subbands
20 the number of computations is approximately one-sixteenth the number required for an equivalent time-domain operation. The phase modification can be ignored if sideways localization is not a necessity or if the computational overhead is viewed excessive.

In step 68, subband data is mixed by multiplying it by the corresponding modified scale factor data and summing it with the scaled subband products of the other eligible
25 subband components in the pipeline (see figure 11). The normal multiplication by step-size, which is dictated by the bit allocation, is avoided by predefining the bit allocation table to be the same for all audio components. The maximum scale factors indexes are looked up and divided into (or multiplied by inverse) the mixed result. The division and multiplication by the inverse operations are mathematically equivalent but the
30 multiplication operation is an order of magnitude faster. Overflow can occur when the mixed result exceeds the value stored in one DWORD. Attempting to store a floating-

point word as an integer creates an exception which is trapped and used to correct the scale factor applied to the affected subband. After the mixing process, data is stored in left shifted form.

5

Assembly and Queuing of Output Data Frames

As shown in figure 4, a controller 70 assembles output frames 72 and places them in a queue for transmission to a surround sound decoder. A decoder will only produce useful output if it can align to the repeating synchronization markers or sync codes embedded within the data stream. The transmission of coded digital audio via a S/PDIF data stream is an amendment of the original IEC958 specification and does not make provision for the identity of the coded audio format. The multiformat decoder must first determine the data format by reliably detecting concurrent sync words and then establish an appropriate decoding method. A loss of sync condition leads to an intermission in the audio reproduction as the decoder mutes its output signal and seeks to re-establish the coded audio format.

Controller 70 prepares a null output template 74 that includes compressed audio representing “silence”. In the currently preferred embodiment, there are no differences in the header information from frame to frame and only the scale factors and subband data regions need to be updated. The template header carries unchanging information regarding the format of the stream bit allocation and the side information for decoding and unpacking information.

Concurrently, the audio renderer is generating the list of sound objects, mapping them to the speaker locations. Within the mapped data, the audible subband data is mixed by the pipeline 82 as described above. The multi-channel subband data generated by the pipelines 82, is compressed (step 78) into FLCs in accordance with the predefined bit allocation table. The pipelines are organized in parallel, each specific to a particular speaker channel.

ITU recommendation BS.775-1 recognizes the limitations of two-channel sound systems for multichannel sound transmissions, HDTV, DVD and other digital audio applications. It recommends three front loudspeakers combined with two rear/side loudspeakers arranged in a constant distance constellation around the listener. For certain

cases where a modified ITU speaker arrangement is adopted then the left surround and right surround channels are delayed 84 by a whole number of compressed audio frames.

A packer 86 packs the scale factor and subband data (step 88) and submits the packed data to controller 70. The possibility of frame overflow is eliminated as the bit allocation tables for each channel in the output stream are predefined. The DTS Interactive format is not bit-rate limited and the simpler and more rapid encoding techniques of linear and block encoding can be applied.

To maintain decoder sync, controller 70 determines whether the next frame of packed data is ready for output (step 92). If the answer is yes, controller 70 writes the packed data (scale factors and subband data) over the previous output frame 72 (step 94) and puts it in the queue (step 96). If the answer is no, controller 70 outputs null output template 74. Sending compressed silence in this manner guarantees the interruption free output of frames to the decoder to maintain sync.

In other words, controller 70 provides a data pump process whose function is to manage the coded audio frame buffers for seamless generation by the output device and without introducing intermissions or gaps in the output stream. The data pump process queues the audio buffer that has most recently completed output. When a buffer finishes output it is reposted back to the output buffer queue and flagged as empty. This empty state flag permits a mixing process to identify and copy data into that unused buffer simultaneously as the next buffer in the queue is output and while the remaining buffers wait for output. To prime the data pump process the queue list must first be populated with null audio buffer events. The content of the initialization buffers whether coded or not should represent silence or other inaudible or intended signal. The number of buffers in the queue and size of each buffer influences the response time to user input. To keep latency low and provide a more realistic interactive experience the output queue is restricted to two buffers in depth while the size of each buffer is determined by the maximum frame size permitted by the destination decoder and by acceptable user latency.

Audio quality may be traded off against user latency. Small frame sizes are burdened by the repeat transmission of header information, which reduces the number of bits available to code audio data thereby degrading the rendered audio while large frame

sizes are limited by the availability of local DSP memory in the home theater decoder thereby increasing user latency. Combined with the sample rate the two quantities determine the maximum refresh interval for updating the compressed audio output buffers. In the DTS interactive system this is the time-base that is used to refresh the 5 localization of sounds and provide the illusion of real-time interactivity. In this system the output frame size is set to 4096 bytes offering a minimum header size, good time resolution for editing and loop creation and low latency to user responses. Typically 69ms to 92ms for a frame size of 4096 bytes and from 34ms to 46ms for a frame size of 10 2048 bytes. At each frame time the distance and angle of an active sound relative to the listener's position is calculated and this information used to render individual sounds. As an example refresh rates of between 31Hz to 47Hz depending on sample rate are possible for a frame size of 4096 bytes.

Looping Compressed Audio

15 Looping is a standard gaming technique in which the same sound bits are looped indefinitely to create a desired audio effect. For example, a small number of frames of a helicopter sound can be stored and looped to produce a helicopter for as long as the game requires. In the time domain, no audible clicking or distortion will be heard during the transition zone between the ending and the starting positions of the sound if the amplitude 20 of the beginning and ends are complementary. This same technique does not work in the compressed audio domain.

Compressed audio is contained in packets of data encoded from fixed frames of 25 PCM samples and further complicated by the inter-dependence of compressed audio frames on previously processed audio. The reconstruction filters in the DTS surround sound decoder delays the output audio such that the first audio samples will exhibit a low level transient behavior due to the properties of the reconstruction filter.

As shown in figure 5, the looping solution implemented in the DTS Interactive system is done off-line to prepare component audio for storage in a compressed format 30 that is compatible with the real-time looping execution in the interactive gaming environment. The first step of the looping solution requires the PCM data of a looped sequence to be first compacted or dilated in time to fit precisely within the boundaries

defined by a whole number of compressed audio frames (step 100). Encoded data is representative of a fixed number of audio samples from each encoded frame. In the DTS system the sample duration is a multiple of 1024 samples. To begin, at least N frames of uncompressed ‘lead-out’ audio are read out from the end of the file (step 102) and temporally appended to the start of the looped segment (step 104). In this example N has value 1 but any value sufficiently large to cover the reconstruction filters dependency on previous frames may be used. After encoding (step 106), N compressed frames are deleted from the beginning of the encoded bit-stream to yield a compressed audio loop sequence (step 108). This process ensures that the values resident in the reconstruction synthesis filter during the closing frames is in agreement with the values necessary to ensure seamless concatenation with the commencing frame and in doing so prevent audible clicking or distortion. On looped playback the read pointers are directed back to the start of the looped sequence for glitch free playback.

15 DTS Interactive Frame Format

A DTS Interactive frame 72 consists of data arranged as shown in figure 6. The header 110 describes the format of the content, the number of subbands, the channel format, sampling frequency and tables (defined in the DTS standard) required to decode the audio payload. This region also contains a sync word to identify the start of the header and provide alignment of the encoded stream for unpacking.

The scale factor section 114 identifies the scale factor for each of the subbands, e.g. 32-subbands. The scale factor data varies from frame-to-frame with the corresponding subband data.

Lastly, the subband data section 116 includes all of the quantized subband data.

5 As shown in figure 7, each frame of subband data consists of 32 samples per subband organized as four vectors 118a – 118d of size eight. Subband samples can be represented by linear codes or by block codes. Linear codes begin with a sign bit followed by the sample data while block codes are efficiently coded groups of subband samples inclusive of sign. The alignment of the bit allocation 112 and scale factors 114 with subband data

10 116 is also depicted.

Subband-Domain Mixing of Compress Audio

As described previously, DTS Interactive mixes the component audio in a compressed format, e.g. subband data, rather than the typical PCM format and thus

15 realizes tremendous computational, flexibility and fidelity benefits. These benefits are obtained by discarding those subbands that are inaudible to the user in two stages. First, the gaming programmer can, based on a priori information about the frequency content of a specific audio component, discard the upper (high frequency) subbands that contain little or no useful information. This is done off-line by setting the upper band bit

20 allocations to zero before the component audio is stored.

More specifically, sample rates of 48.0kHz, 44.1kHz and 32.0kHz are frequently encountered in the audio but the higher sample rates offer high fidelity full bandwidth audio at the cost of memory. This can be wasteful of resources if the material contains little high frequency content such as voice. Lower sample rates may be more appropriate

25 for some material but the problems of mixing differing sample rates arise. Game audio frequently uses the 22.050kHz sampling rate as a good compromise between both audio quality and memory requirements. In the DTS Interactive system all material is encoded at the highest supported sample rate mentioned earlier and material that does not fully occupy the full audio spectrum is treated as follows. Material intended for encoding at

30 say 11.025kHz is sampled at 44.1kHz and the upper 75% of subbands describing the high frequency content are discarded. The result is an encoded file that retains compatibility

and ease of mixing with other higher fidelity signals and yet allows a reduced file size. It is easy to see how this principle can be extended to enable 22.050kHz sampling by discarding the upper 50% of subbands.

Second, DTS Interactive unpacks the scale factors (step 120) and uses them in a simplified psychoacoustic analysis (see figure 9) to determine which of the audio components selected by the map function (step 54) are audible in each subband (step 124). A standard psychoacoustic analysis that takes into account neighboring subbands could be implemented to achieve marginally better performance but would sacrifice speed. Thereafter, the audio renderer unpacks and decompresses only those subbands that are audible (step 126). The renderer mixes the subband data for each subband in the subband domain (step 128), recompresses it and formats it for packing as shown in figure 4 (item 86).

The computational benefits of this process are realized from having to unpack, decompress, mix, recompress and pack only those subbands that are audible. Similarly, because the mixing process automatically discards all of the inaudible data, the gaming programmer is provided greater flexibility to create richer sound environments with a larger number of audio components without raising the quantization noise floor. These are very significant advantages in a real-time interactive environment where user latency is critical and rich high fidelity immersive audio environment is the goal.

20

Psychoacoustic Masking Effects

Psychoacoustic measurements are used to determine perceptually irrelevant information, which is defined as those parts of the audio signal which cannot be heard by human listeners, and can be measured in the time domain, the subband domain, or in some other basis. Two main factors influence the psychoacoustic measurement. One is the frequency dependent absolute threshold of hearing applicable to humans. The other is the masking effect that one sound has on the ability of humans to hear a second sound played simultaneously or even after the first sound. In other words the first sound, in the same or neighboring subband, prevents us from hearing the second sound, and is said to mask it out.

In a subband coder the final outcome of a psychoacoustic calculation is a set of numbers which specify the inaudible level of noise for each subband at that instant. This computation is well known and is incorporated in the MPEG 1 compression standard ISO/IEC DIS 11172 "Information technology - Coding of moving pictures and associated audio for digital storage media up to about 1.5 Mbits/s," 1992. These numbers vary dynamically with the audio signal. The coder attempts to adjust the quantization noise floor in the subbands by way of the bit allocation process so that the quantization noise in these subbands is less than the audible level.

DTS Interactive currently simplifies the normal psychoacoustic masking operation by disabling the inter-subband dependence. In the final analysis, the calculation of the intra-subband masking effects from the scale factors will identify the audible components in each subband, which may or may not be the same from subband to subband. A full psychoacoustic analysis may provide more components in certain subbands and completely discard other subbands, most likely the upper subbands.

As shown in figure 9, the psychoacoustic masking function examines the object list and extracts the maximum modified scale value for each subband of the supplied component streams (step 130). This information is input to the masking function as a reference for the loudest signal that is present in the object list. The maximum scale factors are also directed to the quantizer as the basis for encoding the mixed results into the DTS compressed audio format.

For DTS-domain filtering, the time-domain signal is not available, so masking thresholds are estimated from the subband samples in the DTS signal. A masking threshold is calculated for each subband (step 132) from the maximum scale factor and the human auditory response. The scale factor for each subband is compared to the masking threshold for that band (step 136) and if found to be below the masking threshold set for that band then the subband is considered to be inaudible and removed from the mixing process (step 138) otherwise the subband is deemed to be audible and is kept for the mixing process (step 140). The current process only considers masking effects in the same subband and ignores the effects of neighboring subbands. Although this reduces performance somewhat, the process is simpler and hence much faster as required in an interactive real-time environment.

Bit Manipulation

As discussed above, DTS Interactive is designed to reduce the number of computations required to mix and render the audio signal. Significant effort is expended
5 to minimize the quantity of data that must be unpacked and repacked because these and the decompress/recompress operations are computationally intensive. Still the audible subband data must be unpacked, decompressed, mixed, compressed and repacked. Therefore, DTS Interactive also provides a different approach for manipulating the data to reduce the number of computations to unpack and pack the data as shown in figures
10 10a-10c and to mix the subband data as shown in figure 11.

Digital Surround systems typically encode the bit stream using variable length bit fields to optimize compression. An important element of the unpacking process is the signed extraction of the variable length bit fields. The unpacking procedure is intensive due to the frequency of executing this routine. For example to extract an N-bit field, 32-bit (DWORD) data is first shifted to the left to locate the sign bit in the left most bit field.
15 Next, the value is divided by powers of two or right shifted by (32-N) bit positions to introduce the sign extension. The large number of shifting operations take a finite time to execute and unfortunately cannot be executed in parallel or pipelined with other instructions on the present generation of Pentium processors.

20 DTS Interactive by takes advantage of the fact that the scale factor is related to the bit width size and realizes that this provides the possibility to ignore the final right shifting operation if a) in its place the scale factors are treated accordingly and b) the number of bits that represent the subband data are sufficient that the “noise” represented by the (32-N) right most bits is below the noise floor of the reconstructed signal.
25 Although N may be only a few bits this typically only occurs in the upper subbands where the noise floor is higher. In VLC systems that apply very high compression ratios the noise floor could be exceeded.

As shown in figure 10a, a typical frame will include a section of subband data
140, which includes each piece of N-bit subband data 142 where N is allowed to vary
30 across the subbands but not the samples. As shown in figure 10b, the audio renderer

extracts the section of subband data and stores it in local memory, typically as 32-bit words 144 where the first bit is the sign bit 146 and the next thirty-one bits are data bits.

As shown in figure 10c, the audio renderer has shifted subband data 142 to the left so that its sign bit is aligned with sign bit 146. Since all of the data is stored as FLCs rather than VLCs this is a trivial operation. The audio renderer does NOT right shift the data. Instead, the scale factors are prescaled by dividing them by 2 raised to the power of (32-N) and stored and the 32-N rightmost bits 148 are treated as inaudible noise. In other words, a one bit left shift of the subband data combined with a one bit right shift of the scale factor does not alter the value of the product. The same technique can also be utilized by the decoder.

After summation of all mixing products and quantization it is a simple matter to identify those values that will overflow since the storage limit is fixed. This offers greatly superior detection speed in comparison to a system where the subband data has not been treated by the left shift operation.

When the data is repacked, the audio renderer simply grabs the leftmost N-bits from each 32-bit word thereby avoiding 32-N left shift operations. The avoidance of (32-N) right and left shift operations may seem to be rather insignificant but the frequency of executing the unpack and pack routines is so high that it represents a significant reduction in computations.

20

Mixing Subband Data

As shown in figure 11, the mixing process commences and the audible subband data is multiplied by the corresponding scale factor, which has been adjusted for position, equalization, phase localization etc, (step 150) and the sum is added to the corresponding subband products of the other eligible items in the pipeline (step 152). Since the number of bits for each component in a given subband is the same the step size factors can be ignored thus saving computations. The maximum scale factors indexes are looked up (step 154) and the inverse is multiplied by the mixed result (step 156).

Overflow can occur when the mixed result exceeds the value stored in one DWORD (step 158). Attempting to store a floating point word as an integer creates an exception which is trapped and used to correct the scale factor applied to all affected

subbands. If the exception occurs, the maximum scale factor is incremented (step 160) and the subband data is recalculated (step 156). The maximum scale factors are used as a starting point because it is better to err on the conservative side and increment the scale factor rather than reduce the dynamic range of the signal. After the mixing process, data 5 is stored in left shifted form by modification of the scale factor data for recompression and packing.

While several illustrative embodiments of the invention have been shown and described, numerous variations and alternate embodiments will occur to those skilled in the art. For example, two 5.1 channel signals could be mixed and interleaved together to 10 produce a 10.2 channel signal for true 3D immersion with the added dimension of height. In an addition processing combination, instead of processing one frame at a time, the audio renderer could reduce the frame size by one-half and process two frames at a time. This would reduce latency by one-half but at the cost of wasting some bits on repeating 15 the header information twice as often. However, in a dedicated system much of the header information could be eliminated. Such variations and alternate embodiments are contemplated, and can be made without departing from the spirit and scope of the invention as defined in the appended claims.